CONTENTS

More detailed tables of contents are to be found within the various parts of the compendium. The following provides merely an overview.

ACKNOWLEDGMENTS ii

INTRODUCTION iii

GLOSSARY and LANDMARKS v

PART I. NUCLEIC ACID ALIGNMENTS AND SEQUENCES

Introduction I-1

Contents I-3

A. HIV-1 Alignments and Sequences

Nucleotide Alignments and Consensus Sequences I-A-1

Sequences of WEAU and IBNG I-A-391

HIV-1 Sequence Summary Tables I-A-401

B. HIV-2/SIV Alignments

Nucleotide Alignments and Consensus Sequences I-B-1

HIV-2/SIV Sequence Summary Tables I-B-132

C. AGM Alignments

Nucleotide Alignments and Consensus Sequences I-C-1

AGM Sequence Summary Tables I-C-60

PART II. AMINO ACID ALIGNMENTS

Introduction II-1

Contents II-3

A. HIV-1 Alignments

Amino Acid Alignments and Consensus Sequences II-A-1

B. HIV-2/SIV Alignments

Amino Acid Alignments and Consensus Sequences II-B-1

C. SIVAGM, SIVMND, and SIVSYK Alignments

Amino Acid Alignments and Consensus Sequences II-C-1

PART III ANALYSIS

Contents III-1

HIV Vpr III-2

Host Proteins Associated with HIV-1 . . . . III-10

Sequencing Primers for HIV-1 . . . . III-15

Recombination in HIV-1 and HIV-2 . . . . III-22

Genotyping of HIV-1 . . . . III-30

Scanning for HIV-1 Recombinants III-35

Detection of HIV Hybrids using VESPA III-61

Global Variation in the HIV-1 V3 Region III-77

A New Genetic Subtype of HIV-1 . . . . III-147

PART IV. RELATED SEQUENCES

Introduction and Contents IV-1

Sequence Entries IV-2

PART V. DATABASE COMMUNICATIONS

Introduction to the World Wide Web V-1

References V-2

ACKNOWLEDGMENTS

The HIV Sequence Database and Analysis Project is funded by the Vaccine and Prevention Research Program of the AIDS Division of the National Institute of Allergy and Infectious Diseases (Dr. James Bradac, Project Officer) through an interagency agreement with the U.S. Department of Energy.

We thank the many researchers who have made their sequences available prior to publication.

The photograph on the cover of this compendium of the late Howard Temin was taken by Gregory Anderson and was kindly provided by Bette Sheehan of the University of Wisconsin.

I

Nucleic Acid Alignments and Sequences

Nucleotide sequence alignments were generated using the PIMA program developed by Randy Smith and Temple Smith (*Protein Engineering* **5**:35, 1992). By this approach, a protein pattern is created which then serves as a template for the nucleotide sequence alignment. Hand-editing, when required, was done with the MASE program (D. V. Faulkner and J. Jurka, *TIBS* **13**:321, 1988). For information concerning either of these programs, contact Dr. Randall Smith, Institute of Molecular Genetics, Baylor College of Medicine, Houston, Texas, 713-798-4735.

With few exceptions, only full-length coding sequences were included in these alignments. Tables of information pertaining to each sequence in an alignment are provided for the first time in 1995. The common names given to sequences in alignment and in the accompanying tables were selected on several grounds: for sequences corresponding to samples provided by the NIAID repository, WHO and DAIDS conventions are employed for the names; for other sequences, the name given by the authors of the paper reporting the sequence are usually utilized. Locus names, which no longer appear in alignments but are provided in the tables, will usually correspond to GenBank Locus names, however there is no guarantee for that and in any case EMBL and DDBJ names may differ. The Accession numbers, also provided in the tables, will be universal identifiers. We wish to thank many sequencers for providing data prior to publication.

Mixed case consensus sequences are used as the reference sequences for each alignment. Upper case letters indicate 100% conservation of nucleotide bases in a given position of the alignment, and lower case letters represent bases conserved in at least 50% of the sequences. The symbol ”?” indicates no consensus at a position. To reduce the alignment uncertainty with the large HIV-1 set, sequence subtypes (see below and Part III) were constituted and consensus sequences were generated for each. Divergent sequences that could not be easily assigned to any given subtype and mosaic sequences (perhaps the result of recombinational events) are shown at the bottom of each block. Large deletions, if present in only one sequence (*e.g.*, the HIV2nihz *nef* coding region sequence), were also excluded from the consensus sequences.

Following the alignment for each coding sequence, just the consensus sequences over that coding region are re-presented in alignment.

Prior to the 1992 database compendium, HIV-1 sequences had been roughly classified as ‘U.S.’ and ‘African.’ In light of many new HIV-1 *gag* and *env* sequences, it became more useful, starting with the 1992 compendium, to categorize HIV-1 sequences into five sequence subtypes, depending upon the coding sequence. HIV-1 subtypes now number eight (A through H). Collectively, these are called group M sequences, as they are significantly distinct from group O sequences. The bases for this classification, discussed at greater length in Part III of the 1994 compendium, are:

i) subtypes are approximately equidistant from one another in *env* (a “star phylogeny”);

ii) the *env* phylogenetic tree is for the most part congruent with *gag* phylogenetic trees;

iii) two or more samples are required to define a sequence subtype.

Subtype naming problems have arisen for several reasons. A small but not insignificant number of viral sequences are hybrid, clustering with one sequence subtype in *gag* and another sequence subtype in *env*, for example; or, to take another example, clustering over different stretches with two or more subtypes in *env*. All subtype E sequences based on *env* have *gag* sequences that align with subtype A sequences; and, moreover, the 3’ half of the gp41 cds of E and subtype G *env* sequences align with subtype A sequences (Gao *et al*, J. Virol., in press, 1996). Given the homogeneity of G and E subtype sequences, these are handled as subtypes and not as hybrids in the following alignments. It remains to be discovered whether these states have arisen from recombination or lack of divergence.

Several analyses of mosaic molecules are presented in Part III. Naming also becomes problematic when highly divergent forms of a given subtype arise: such forms are sometimes designated A', B', F', etc. It is increasingly necessary to have sequence data from both *gag* and *env* coding sequences when a new form or subtype is being claimed.

There are at least five sequence subtypes for the HIV-2s, A through E.

The authority for some of the annotation is limited largely to invariance—the recurrence of patterns such as AATAAA, for example. The reader should be cautious in drawing upon this information. Due to space limitations, only certain sequences are reported as entries in this section. Beginning in 1995, most sequences are cataloged in Part I using just the headers of GenBank entries; to gain the full entry including the sequence itself, users should go to the HIV Sequence Database WWW site (*http://hiv-web.lanl.gov*), the FTP site (also described in Part V of this compendium) or any of the large gene libraries. Sequences used in alignments are provided on the accompanying diskettes, and the alignments themselves are on the WWW site. Sequence entries from previous years are presented in the 1990–1994 compendiums.